International Journal of Agriculture
Science and Research (IJASR)
ISSN (P): 2250–0057; ISSN (E): 2321–0087
Vol. 11, Issue 2, Dec 2021, 15-22
© TJPRC Pvt. Ltd.

•Journal Publications •Research Consultancy

Original Article

REVIEW OF DATA MINING CLASSIFICATION TECHNIQUES

SHRADDHA SHARMA & ANKITA SAXENA

Department of Computer Science & Engineering, Prestige Institute of Engineering Management & Research, Indore

ABSTRACT

One of the most valuable tools for scholars, individuals, and developers is data mining, which allows them to obtain relevant knowledge from large amounts of data. Data mining is the process of using data collection techniques to discover real correlations and associations in large data sets that are obscure. Data Mining is also known as Information Discovery of Databases Data classification, data integration, data collection, data mining, data selection, knowledge presentation, and pattern evaluation are all part of knowledge discovery. Data classification is a crucial method for dealing with vast amounts of data. It is used to assign a class mark to newly available records. We find a paradigm of Classification that defines data classes and differentiates them from other classes. Building accurate and fast classifiers for large data sets is a crucial challenge in data mining and knowledge exploration. We use classification based on the training set to predict class labels and to classify results. This paper provides a detailed description of data mining and information discovery labelling strategies. In addition, this paper examines a number of approaches, including K-Nearest Neighbour, Bayesian Classifiers, Decision Tree, Genetic Algorithm, Fuzzy Set Approach, and Neural Networks. The aim of this research is to provide a concise overview of various data mining classification techniques.

KEYWORDS: - Data Mining, Classification, Categories, Class Level & Prediction, Model

Received: May 18, 2021; **Accepted**: Jun 08, 2021; **Published**: Jun 09, 2021; **Paper Id.**: IJASRDEC20212

1. INTRODUCTION

Data mining is the method of collecting valuable information from vast volumes of data. To put it another way, it's the process of extracting information from data. Huge sets of data are first sorted, then patterns are found and relationships are formed to conduct data analysis and solve problems in the data mining process. Researchers have recently developed and used a variety of Data Mining strategies, including classification, grouping, clustering, regression, intra, temporal patterns, prediction, and associations. On the basis of a training collection of data comprising findings, classification implies determining which set of groups a new observation belongs to. The data classification process includes two steps-

- Building the classifier or model
- Using classifier for classification

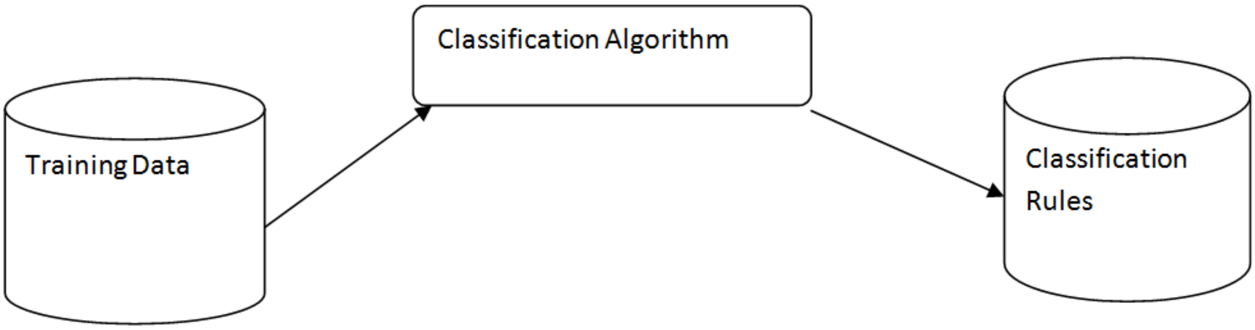

Figure 1: Construction of Classifier

Step 1: Building the classifier: This is the stage of studying or teaching. To construct a classifier, different classification algorithms are used in this process. Server tuples and their respective class labels make up the training list. To build a classifier, this training set is used. These tuples are also known as objects, samples, and data points. A division or class is the name given to each tuple that makes up the training set.

Step 2: Using classifier for classification: The classifier model is used to approximate the consistency of the classification rules by predicting the reasoning and evaluating the constructed model on test results. Fresh data tuples are subjected to the guidelines to see if the precision is sufficient.

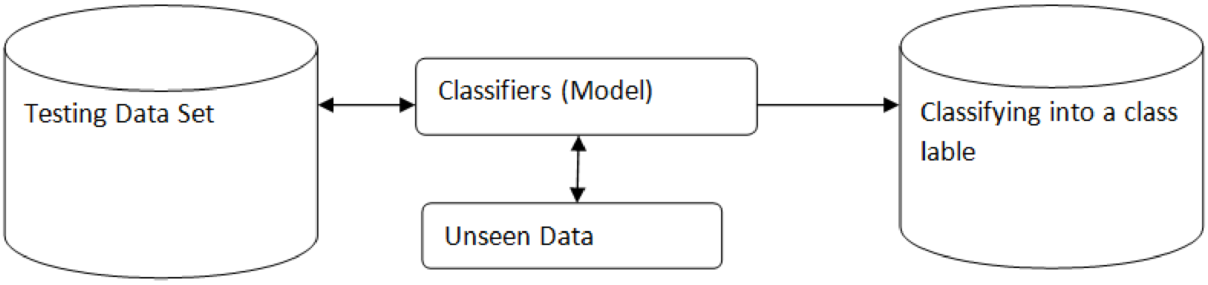

Figure 2: Use of Classifier

2. Issues with Classification of Data

The below are some of the issues with data preparation for classification:

- Data cleaning entails the elimination of excess noise as well as the treatment of missed values. Noise removal systems are used to eliminate the noise. The problem of missing values is overcome by substituting the most frequently occurring value for the attribute in place of the missing value.
- Data transformation and reduction: There are two methods for transforming data. The first step is to scale all of the values for a given attribute. The data is converted in the second generalisation by generalising it to a higher definition.
- Relevance Analysis: It's possible that the database contains characteristics that aren't important. Correlation analysis is used to determine whether two attributes are related.

3. Characteristics of Classifiers

Any classifier has a feature that distinguishes it from other classifiers. These characteristics are referred to as classifier characteristics. The following are some of these characteristics:

Accuracy refers to a classifier's capacity. This property determines how a classifier correctly classifies a data set. It accurately predicts the class name.

Speed: The computational cost of creating and using a classifier is referred to as the classifier's speed. The time it takes to construct a model that can classify the number of tuples.

Robustness: Robustness refers to a classifier's ability to make accurate predictions from a given set of noisy data. It requires correctly identifying tuple data in the presence of noise.

Scalability: The classifier should be designed in such a way that it is unaffected by the scale of the database. Scalability refers to the ability to construct a classifier effectively with a vast volume of data.

Interpretability: It refers to how well the classifier understands the data.

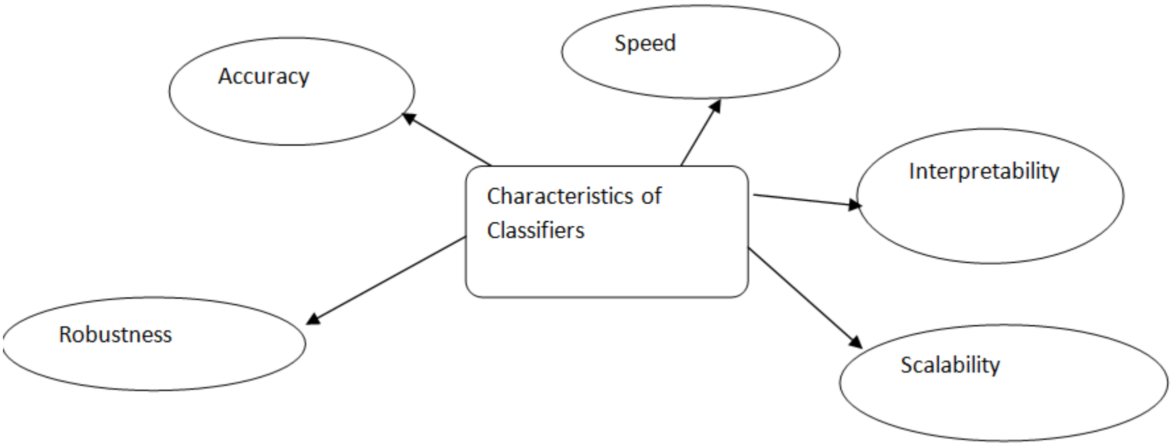

Figure 3: Characteristics of Classifiers

4. RELATED WORK

Classification has become an increasingly important application in recent years, because of its unique ability to enable and distinguishes data from other classes. An increased level of interest in the field of social networking has also resulted in a revival of classification algorithms. Therefore, a number of techniques have recently been designed for a better understanding of huge data set.

No.	Year	Algorithm	Merits
1	2012	Associative Classification and a Genetic Algorithm are used to forecast heart disease.	Discovered rules are highly comprehensible. Best prediction of heart disease
2	2013	Artificial Neural Network classification of heart disease and function subset collection	Eliminates useless and distortive data
3	2014	For medical diagnostic interface, a fuzzy weighted association rule mining classifier dependent on the gain ratio	improve the classifier accuracy
4	2015	Data Mining Classification Algorithm Accuracy and Training Times	For various data sizes, the time taken by the algorithms for training and the accuracy of their classifications were investigated.
5	2015	Cancer Diagnosis and Prognosis Using Data Mining Techniques	the patterns frequently found in benign and malignant patients
6	2016	Data mining Classification Technique for Appraisal Management System	predict the potential talent that for promotion or no
7	2016	Data Mining Classification Techniques' Performance in Public Health Care	The best methodology for a given data set is selected based on the highest level of precision.

The algorithm "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" was proposed by Akhil Jabbar et al. in 2012. It is a powerful classification algorithm that uses a genetic approach to forecast heart disease. The key gain is the discovery of highly comprehensible multi-level estimation rules with high predictive precision and high interestingness values. The proposed algorithm aids in the most accurate prediction of heart disease, as well as physicians' diagnostic decisions.[1].

Akhil Jabbar et al. proposed an algorithm in 2013. It is a modern function selection tool for heart disease classification that employs ANN. They used various feature selection approaches to rate the features that add the most to the assessment of heart disease, thus reducing the number of diagnostic tests that a patient would undergo. The proposed approach avoids redundant and erroneous data. [2].

Previous algorithms focused on knowledge gain and fuzzy association rule mining, as shown by N. S. Nithya et al. in 2014, are not feasible. Nithya changed the benefit ratio based fuzzy weighted association rule mining by using a vast number of distinct values. This algorithm improves the classifier's accuracy.[3].

S. Olalekan Akinola and O. Jephthar Oyabugbe demonstrated how data mining classification algorithms work as input data sizes grow in 2015. They used three classification algorithms based on data mining. Various simulated data sizes were used to test the Multi-Layer Perceptron (MLP) Neural Network, Decision Tree, and Nave Bayes. For various data sizes, the time taken by the algorithms for training and the accuracy of their classifications were investigated.[4].

Nikhil N. Salvithal and R.B. Kulkarni proposed various classifier algorithms on the Talent dataset in 2016 to spot the talent collection and judge the individual's results. Finally, based on accuracy, the best-suited classifier is selected, and this system is used to create classification rules to predict whether or not a possible talent is suitable for promotion.[6].

Tanvi Sharma and Anand Sharma proposed an algorithm in 2016 that relies on the use of multiple data mining classification techniques and data mining methods to analyse the health-care system. As an output metric, the percentage of accuracy of each applied data mining classification technique is used. The best methodology for a given data set is selected based on the highest level of precision.[7].

5. CLASSIFICATION MODELS

The key objectives of a classification algorithm are to improve the classification model's predictive accuracy. For classification, a variety of model strategies are used, including the following: [8,9,10].

5.1 Decision Tree

Decision tree build classification model in form of tree, which is a hierarchical data structure consisting of node and directed edges. A tree has three type of node i.e. root node, internal nodes contain attribute test condition and leaf or terminal nodes is assigned a class label. It uses an if-them rule set.

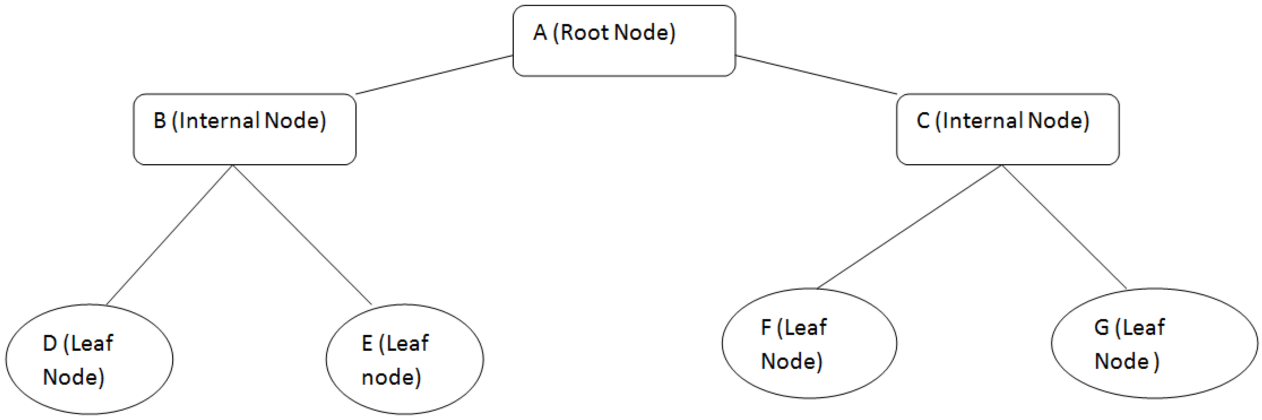

Figure 4: Decision Tree

5.2 Bayesian Classifiers

Probabilistic classifiers are Bayesian classifiers. Each tuple has an n-dimensional attribute set, A1, A2,..., An. P(X) is stable for all classes; the only thing that has to be maximised is P (X/Ci) P (Ci). If the prior probabilities for the groups are unknown, it is normal to conclude that they are similarly probable, i.e., P(C1) = P(C2) =...=P(Cm), and we will optimise P(X/Ci). Otherwise, P(X/Ci)P is maximised (Ci).

$$P(\mathbf{X} | C_i) = \prod_{k=1}^{n} P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \ldots \times P(x_n | C_i)$$

5.3 k- Nearest Neighbor

In an n-dimensional space, the k-Nearest Neighbor algorithm stores all instances corresponding to training data points. A k-nearest neighbour classifier saves the closest k number of instances when an unknown sample is obtained. For a given unknown value, it returns the most common class as a forecast, and for real-valued results, it returns k nearest neighbour. It uses the following formula to weigh the contributions of each of the k neighbours based on their distance:

$$w \equiv \frac{1}{d(x_q, x_i)^2}$$

5.4 Genetic Algorithm

The initial population in a genetic algorithm is made up of randomly generated rules. Each rule can be interpreted as a string of bits. The survival of the fittest principle underpins the genetic algorithm. The consistency of a rule's classification on a collection of given samples determines its fitness. Following that, crossover and mutation are included. Finally, offspring are born.

5.5 Neural Network

Pattern analysis and classification was done using neural networks. It's a set of input and output units with a weight assigned to each attachment. The network learns by changing the weights until it can correctly predict the class mark of input tuples in the first step, known as the learning phase.

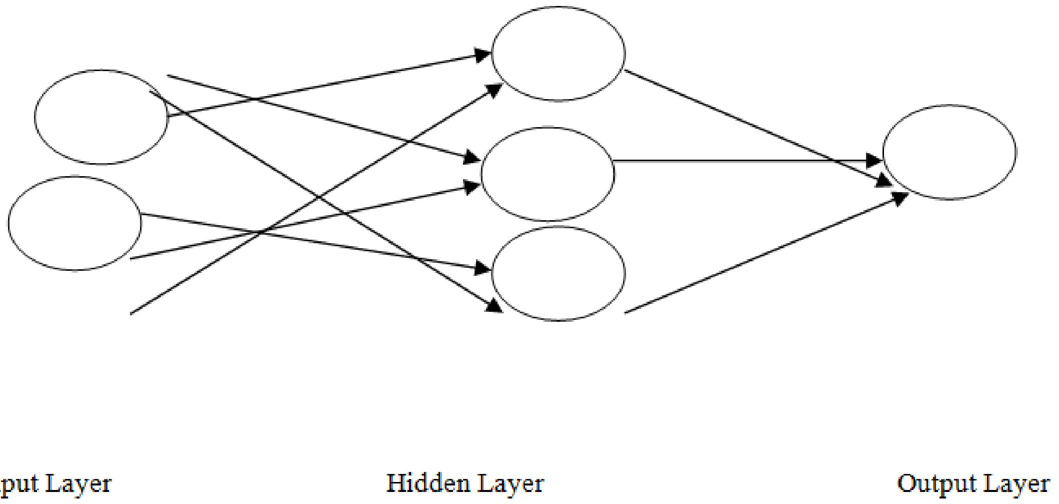

Figure 5: Neural Network

Artificial Neural Networks, on the other hand, have worked admirably in the majority of real-world implementations. It has a good tolerance for noisy data and can classify patterns that haven't been taught. Artificial Neural Networks usually do best where the inputs and outputs are continuous.

5.6. Fuzzy Set Approaches

Possibility Theory is another name for Fuzzy Set. It enables one to function at a high abstraction level. Inexact evidence should be dealt with using fuzzy set theory. In fuzzy set theory, an entity may be assigned to several fuzzy sets.

6. Advantages and Disadvantages

S.No.	Advantage	Disadvantage
1.	Methods are cost effective and efficient	Privacy Issues
2.	Identify Criminal Suspects	Accuracy problem
3.	Predicting risk of diseases	Assumption is independence of features
4.	Helps bank and Financial institutions to identify defaulters	Need to determine values of parameter
5.	Simple to comprehend and justify	For uncorrelated variables, this method does not function well.
6.	Handle real and discrete data	Information extraction is a complex task.

7. CONCLUSIONS

In data processing, there are a number of classification strategies. Per strategy has its own set of benefits and drawbacks. Some strategies, such as Decision tree classifiers, Bayesian classifiers, and classification by back propagation, are quick learners that use training tuples to build a generalization model. Lazy learner classification algorithms include k-nearest-neighbor classifiers and case-based inference.

REFERENCES

1. *M. Akhil Jabbar & Dr. Priti Chandrab "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.*
2. *M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013 International Research Journal Publisher: Global Journals Inc. (USA)*
3. *N S Nithyaand K Duraiswamy "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences*
4. *S. Olalekan Akinola, O. Jephthar Oyabugbe Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" Journal of Software Engineering and Applications, 2015, 8, 470-477 Published Online September 2015 in SciRes. http://www.scirp.org/journal/jsea*
5. *Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015*
6. *Nikhil N. Salvithal " Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016*
7. *Tanvi Sharma, Anand Sharma & Vibhakar Mansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016*
8. *B Rosiline Jeetha "EFFICIENT CLASSIFICATION METHOD FOR LARGE DATASET BY ASSIGNING THE KEY VALUE IN CLUSTERING" International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IJCSMC, Vol. 3, Issue. 1, January 2014, pg.319 – 324*
9. *Divya Tomar and Sonali Agarwal " A survey on Data Mining approaches for Healthcare" international Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25*
10. *V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 – 45*